Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume II
Data Pro–I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Instance Selection

Huan Liu
Arizona State University, USA

Lei Yu
Arizona State University, USA

INTRODUCTION

The amounts of data become increasingly large in recent years as the capacity of digital data storage worldwide has significantly increased. As the size of data grows, the demand for data reduction increases for effective data mining. Instance selection is one of the effective means to data reduction. This article introduces basic concepts of instance selection, its context, necessity and functionality. It briefly reviews the state-of-the-art methods for instance selection.

Selection is a necessity in the world surrounding us. It stems from the sheer fact of limited resources. No exception for data mining. Many factors give rise to data selection: data is not purely collected for data mining or for one particular application; there are missing data, redundant data, and errors during collection and storage; and data can be too overwhelming to handle. Instance selection is one effective approach to data selection. It is a process of choosing a subset of data to achieve the original purpose of a data mining application. The ideal outcome of instance selection is a model independent, minimum sample of data that can accomplish tasks with little or no performance deterioration.

BACKGROUND AND MOTIVATION

When we are able to gather as much data as we wish, a natural question is "how do we efficiently use it to our advantage?" Raw data is rarely of direct use and manual analysis simply cannot keep pace with the fast accumulation of massive data. Knowledge discovery and data mining (KDD), an emerging field comprising disciplines such as databases, statistics, machine learning, comes to the rescue. KDD aims to turn raw data into nuggets and create special edges in this ever competitive world for science discovery and business intelligence. The KDD process is defined (Fayyad *et al.*, 1996) as *the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*. It includes data selection, preprocessing, data mining, interpretation and evaluation. The first two processes (data selection and preprocessing) play a pivotal role in successful data mining (Han and Kamber, 2001). Facing the mounting challenges of enormous amounts of data, much of the current research concerns itself with scaling up data mining algorithms (Provost and Kolluri, 1999). Researchers have also worked on scaling down the data - an alternative to the algorithm scaling-up. The major issue of scaling down data is to select the relevant data and then present it to a data mining algorithm. This line of work is in parallel with the work on algorithm scaling-up and the combination of the two is a two-edged sword in mining nuggets from massive data.

In data mining, data is stored in a *flat file* and described by terms called *attributes* or *features*. Each line in the file consists of attribute-values and forms an *instance*, also named as a *record, tuple*, or *data point* in a multi-dimensional space defined by the attributes. Data reduction can be achieved in many ways (Liu and Motoda, 1998; Blum and Langley, 1997; Liu and Motoda, 2001). By selecting features, we reduce the number of columns in a data set; by discretizing feature-values, we reduce the number of possible values of features; and by selecting instances, we reduce the number of rows in a data set. We focus on instance selection here.

Instance selection reduces data and enables a data mining algorithm to function and work effectively with huge data. The data can include almost everything related to a domain (recall that data is not solely collected for data mining), but one application is normally about using one aspect of the domain. It is natural and sensible to focus on the relevant part of the data for the application so that search is more focused and mining is more efficient. It is often required to clean data before mining. By selecting relevant instances, we can

usually remove irrelevant, noise, and redundant data. The high quality data will lead to high quality results and reduced costs for data mining.

MAJOR LINES OF RESEARCH AND DEVELOPMENT

A spontaneous response to the challenge of instance selection is, without fail, some form of sampling. Although it is an important part of instance selection, there are other approaches that do not rely on sampling, but resort to search or take advantage of data mining algorithms. In the following, we start with sampling methods, and proceed to other instance selection methods associated with data mining tasks such as classification and clustering.

Sampling Methods

Sampling methods are useful tools for instance selection (Gu, Hu, and Liu, 2001).

$\binom{N}{n}$ *Simple random sampling* is a method of selecting n instances out of the N such that every one of the distinct samples has an equal chance of being drawn. If an instance that has been drawn is removed from the data set for all subsequent draws, the method is called random sampling without replacement. Random sampling with replacement is entirely feasible: at any draw, all N instances of the data set are given an equal chance of being drawn, no matter how often they have already been drawn.

Stratified random sampling The data set of N instances is first divided into subsets of N_1, N_2,…, N_l instances, respectively. These subsets are non-overlapping, and together they comprise the whole data set (i.e., $N_1+N_2,\ldots,+N_l=N$). The subsets are called strata. When the strata have been determined, a sample is drawn from each stratum, the drawings being made independently in different strata. If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. It is often used in applications that we wish to divide a heterogeneous data set into subsets, each of which is internally homogeneous.

Adaptive sampling refers to a sampling procedure that selects instances depending on results obtained from the sample. The primary purpose of adaptive sampling is to take advantage of data characteristics in order to obtain more precise estimates. It takes advantage of the result of preliminary mining for more effective sampling and vise versa.

Selective sampling is another way of exploiting data characteristics to obtain more precise estimates in sampling. All instances are first divided into partitions according to some homogeneity criterion, and then random sampling is performed to select instances from each partition. Since instances in each partition are more similar to each other than instances in other partitions, the resulting sample is more representative than a randomly generated one. Recent methods can be found in (Liu, Motoda, and Yu, 2002) in which samples selected from partitions based on data variance result in better performance than samples selected from random sampling.

Methods for Labeled Data

One key data mining application is classification – predicting the class of an unseen instance. The data for this type of application is usually labeled with class values. Instance selection in the context of classification has been attempted by researchers according to the classifiers being built. We include below five types of selected instances.

Critical points are the points that matter the most to a classifier. The issue was originated from the learning method of Nearest Neighbor (NN) (Cover and Thomas, 1991). NN usually does not learn during the training phase. Only when it is required to classify a new sample does NN search the data to find the nearest neighbor for the new sample and use the class label of the nearest neighbor to predict the class label of the new sample. During this phase, NN could be very slow if the data is large and be extremely sensitive to noise. Therefore, many suggestions have been made to keep only the critical points so that noisy ones are removed as well as the data set is reduced. Examples can be found in (Yu *et al*., 2001) and (Zeng, Xing, and Zhou, 2003) in which critical data points are selected to improve the performance of collaborative filtering.

Boundary points are the instances that lie on borders between classes. Support vector machines (SVM) provide a principled way of finding these points through minimizing structural risk (Burges, 1998). Using a non-linear function φ to map data points to a high-dimensional feature space, a non-linearly separable

data set becomes linearly separable. Data points on the boundaries, which maximize the margin band, are the support vectors. Support vectors are instances in the original data sets, and contain all the information a given classifier needs for constructing the decision function. Boundary points and critical points are different in the ways how they are found.

Prototypes are representatives of groups of instances via averaging (Chang, 1974). A prototype that represents the typicality of a class is used in characterizing a class, instead of describing the differences between classes. Therefore, they are different from critical points or boundary points.

Tree based sampling Decision trees (Quinlan, 1993) are a commonly used classification tool in data mining and machine learning. Instance selection can be done via the decision tree built. In (Breiman and Friedman, 1984), they propose *delegate sampling*. The basic idea is to construct a decision tree such that instances at the leaves of the tree are approximately uniformly distributed. Delegate sampling then samples instances from the leaves in inverse proportion to the density at the leaf and assigns weights to the sampled points that are proportional to the leaf density.

Instance labeling In real world applications, although large amounts of data are potentially available, the majority of data are not labeled. Manually labeling the data is a labor intensive and costly process. Researchers investigate whether experts can be asked to only label a small portion of the data that is most relevant to the task if it is too expensive and time consuming to label all data. Usually an expert can be engaged to label a small portion of the selected data at various stages. So we wish to select as little data as possible at each stage, and use an adaptive algorithm to guess what else should be selected for labeling in the next stage. Instance labeling is closely associated with adaptive sampling, clustering, and active learning.

Methods for Unlabeled Data

When data is unlabeled, methods for labeled data cannot be directly applied to instance selection. The widespread use of computers results in huge amounts of data stored without labels (web pages, transaction data, newspaper articles, email messages) (Baeza-Yates and Ribeiro-Neto, 1999). Clustering is one approach to finding regularities from unlabeled data. We discuss three types of selected instances here.

Prototypes are pseudo data points generated from the formed clusters. The idea is that after the clusters are formed, one may just keep the prototypes of the clusters and discard the rest data points. The *k*-means clustering algorithm is a good example of this sort. Given a data set and a constant *k*, the *k*-means clustering algorithm is to partition the data into *k* subsets such that instances in each subset are similar under some measure. The *k* means are iteratively updated until a stopping criterion is satisfied. The prototypes in this case are the *k* means.

Prototypes plus sufficient statistics In (Bradley, Fayyad, and Reina, 1998), they extend the *k*-means algorithm to perform clustering in one scan of the data. By keeping some points that defy compression plus some sufficient statistics, they demonstrate a scalable *k*-means algorithm. From the viewpoint of instance selection, it is a method of representing a cluster using both defiant points and pseudo points that can be reconstructed from sufficient statistics, instead of keeping only the *k* means.

Squashed data are some pseudo data points generated from the original data. In this aspect, they are similar to prototypes as both may or may not be in the original data set. Squashed data points are different from prototypes in that each pseudo data point has a weight and the sum of the weights is equal to the number of instances in the original data set. Presently two ways of obtaining squashed data are (1) model free (DuMouchel *et al*., 1999) and (2) model dependent (or likelihood based (Madigan *et al*., 2002)).

FUTURE TRENDS

As shown above, instance selection has been studied and employed in various tasks (sampling, classification, and clustering). Each task is very unique in itself as each task has different information available and requirements. It is clear that a universal model of instance selection is out of the question. This short article provides some starting points that can hopefully lead to *more concerted study and development of new methods for instance selection*. Instance selection deals with scaling down data. When we understand better instance selection, it is natural to investigate if *this work can be combined with other lines of research* in overcoming the problem of huge amounts of data, such as algorithm scaling-up, feature selection and construction. It is a big challenge

to integrate these different techniques in achieving the common goal - effective and efficient data mining.

CONCLUSION

With the constraints imposed by computer memory and mining algorithms, we experience selection pressures more than ever. The central point of instance selection is *approximation*. Our task is to achieve as good mining results as possible by approximating the whole data with the selected instances and hope to do better in data mining with instance selection as it is possible to remove noisy and irrelevant data in the process. In this short article, we have presented an initial attempt to review and categorize the methods of instance selection in terms of sampling, classification, and clustering.

REFERENCES

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley and ACM Press.

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, 245–271.

Bradley, P., Fayyad, U., & Reina, C. (1998). Scaling clustering algorithms to large databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 9 – 15.

Burges, C. (1998). A tutorial on support vector machines. *Journal of Data Mining and Knowledge Discovery*, 2,121-167.

Chang, C. (1974). Finding prototypes for nearest neighbor classifiers. *IEEE Transactions on Computers*, C-23.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C., & Pregibon, D. (1999). Squashing flat files flatter. In *Proceedings of the Fifth ACM Conference on Knowledge Discovery and Data Mining*, pp. 6-15.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). From data mining to knowledge discovery. In *Advances in Knowledge Discovery and Data Mining*.

Gu, B., Hu, F., & Liu, H. (2001). Sampling: knowing whole from its part. In *Instance Selection and Construction for Data Mining*. Boston: Kluwer Academic Publishers.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Liu, H. & Motoda, H., (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers.

Liu, H. & Motoda, H., editors (2001). *Instance Selection and Construction for Data Mining*. Boston: Kluwer Academic Publishers.

Liu, H., Motoda, H., & Yu, L. (2002). Feature selection with selective sampling. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 395-402, 2002.

Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C., & Ridgeway, G. (2002). Likelihood-based data squashing: a modeling approach to instance construction. *Journal of Data Mining and Knowledge Discovery,* 6(2), 173-190.

Provost, F. & Kolluri, V. (1999). A survey of methods for scaling up inductive algorithms. *Journal of Data Mining and Knowledge Discovery*, 3, 131 – 169.

Quinlan, R.J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Yu, K., Xu, X., Ester, M., & Kriegel, H. (2001) Selecting relevant instances for efficient and accurate collaborative filtering. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pp. 239-46.

Zeng, C., Xing, C., & Zhou, L. (2003). Similarity measure and instance selection for collaborative filtering. In *Proceedings of the Twelfth International Conference on World Wide Web*, pp. 652-658.

KEY TERMS

Classification: A process of predicting the classes of unseen instances based on patterns learned from available instances with predefined classes.

Clustering: A process of grouping instances into clusters so that instances are similar to one another within a cluster but dissimilar to instances in other clusters.

Data Mining: The application of analytical methods and tools to data for the purpose of discovering patterns, statistical or predictive models, and relationships among massive data.

Data Reduction: A process of removing irrelevant information from data by reducing the number of features, instances, or values of the data.

Instance: A vector of attribute-values in a multi-dimensional space defined by the attributes, also named as a record, tuple, or data point.

Instance Selection: A process of choosing a subset of data to achieve the original purpose of a data mining application as if the whole data is used.

Sampling: A procedure that draws a sample S_i by a random process in which each S_i receives its appropriate probability P_i of being selected.

This work was previously published in Encyclopedia of Data Warehousing and Mining, edited by J. Wang, pp. 621-624, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).